MATHEMATICS OF OPERATIONS RESEARCH
Vol. 9, No. 1, February 1984
Printed in U.S.A.

ON RANDOM BINARY TREES*

GERALD G. BROWN AND BRUNO O. SHUBERT

Naval Postgraduate School

A widely used class of binary trees is studied in order to provide information useful in evaluating algorithms based on this storage structure. A closed form counting formula for the number of binary trees with n nodes and height k is developed and restated as a recursion more useful computationally. A generating function for the number of nodes given height is developed and used to find the asymptotic distribution of binary trees. An asymptotic probability distribution for height given the number of nodes is derived based on equally likely binary trees. This is compared with a similar result for general trees.

Random binary trees (those resulting from a binary tree sorting algorithm applied to random strings of symbols) are counted in terms of the mapping of permutations of n symbols to binary trees of height k. An explicit formula for this number is given with an equivalent recursive definition for computational use. A generating function is derived for the number of symbols given height. Lower and upper bounds on random binary tree height are developed and shown to approach one another asymptotically as a function of n, providing a limiting expression for the expected height.

The random binary trees are examined further to provide expressions for the expectations of the number of vacancies at each level, the distribution of vacancies over all levels, the comparisons required for insertion of a new random symbol, the fraction of nodes occupied at a particular level, the number of leaves, the number of single vacancies at each level, and the number of twin vacancies at each level. A random process is defined for the number of symbols required to grow a tree exceeding any given height.

Finally, an appendix is given with sample tabulations and figures of the distributions.

1. Introduction. A binary tree is a finite set of nodes, either empty or containing one node called a root, such that all other nodes are partitioned into disjoint sets which are respectively called left and right subtrees of the root. The subtrees also satisfy the definition of a binary tree. Thus, a binary tree is an unlabelled rooted arborescence with successors of at most degree two distinguished only as left and right.

Figure 1.1 shows a binary tree with six nodes. The root node is shown at the top and is connected by arcs to two immediate successor nodes which are the roots of its left and right subtrees. Each node with no successors (for instance, the left subtree of the root) is called a leaf. The level of a node indicates how deep it is within the tree. Thus the root has level one, its immediate successor nodes have level two, and so forth down the occupied portions of the subtrees. The height of a binary tree is the largest

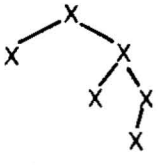

FIGURE 1.1. A binary tree with six nodes.

*Received October 20, 1980; revised July 16, 1982.
AMS 1980 subject classification. Primary: 05C05; Secondary: 68E10.
OR/MS Index 1978 subject classification. Primary: 432 Mathematics/combinatorics. Secondary: 63 Computers/file systems.
Key words: Binary trees, reinformation retrieval, data storage.

0364-765X/84/0901/0043$01.25

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
FEB 1984		**00-00-1984 to 00-00-1984**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
On Random Binary Trees	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Operations Research Department,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**23**	

occupied level. A full binary tree has no internal vacancies (unoccupied node positions).

Binary trees are frequently used as information storage structures on digital computers. For instance, one of the most popular methods of randomly retrieving information by a key, or symbol, is to store the key data in a binary tree. To search for a particular symbol, we begin by looking at the root and proceed by applying the following rules recursively:

(1) If the symbol matches the root symbol, the symbol is found.

(2) If the symbol is "less than" the root (according to some binary ordering relation) continue the search by considering the left successor of the root as the new root (of the left subtree).

(3) If the symbol is greater than the root, continue by searching the right subtree.

(4) If there is no root, the symbol is not in the binary tree.

We assume for simplicity that all symbols are distinct with respect to the ordering relation. Otherwise, the ordering relation and search must be modified in an obvious fashion. The construction of a binary tree for use by such a search scheme may be performed by sequentially examining the key symbols to be inserted. This binary tree sort is a one-pass ordering procedure which proceeds:

(1) If there is no root, insert the symbol as the root.

(2) If the symbol is less than the root symbol, continue by considering the left subtree.

(3) If the symbol is greater than the root symbol, continue with the right subtree.

As an example, consider the six symbols *ABCDEF* and a lexicographical binary ordering relation. Suppose that the particular permutation of symbols examined in *BDAFCE*. The resulting binary tree is shown in Figure 1.2, and has structure identical to the tree in Figure 1.1. Note that this same tree may have resulted from other permutations of the same symbols, for instance *BADCFE*. Therefore, there is a many-to-one mapping of key symbol permutations to corresponding binary trees.

The height of the binary tree in Figure 1.2 is four and thus the maximum number of comparisons required to insert another symbol is four. Similarly, if this tree is used for retrieving symbols, the maximum search length for a symbol in the tree is four, and for a symbol not found the maximum is five.

A computer implementation of a binary tree storage structure requires that each node be represented by its key symbol accompanied by sufficient additional information to identify and access the left and right subtrees. This is usually accomplished by use of a dense array of node symbols each with left and right pointers, by node storage via address calculation into an array space sufficient to store all possible binary trees with a given number of nodes and some maximum height, or by some similar method.

In the following sections we study this widely used class of binary trees in order to provide information useful in examining algorithms based on this storage structure. A closed form counting formula for the number of binary trees with n nodes and height k is developed and restated as a recursion more useful computationally. A generating function for the number of nodes given height is developed and used to find the asymptotic distribution of binary trees. An asymptotic probability distribution for height given the number of nodes is derived based on equally likely binary trees. This is compared with a similar result for general trees.

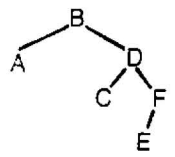

FIGURE 1.2. A binary tree with inserted symbols.

Random binary trees (those resulting from the binary tree sorting algorithm applied to random strings of symbols) are counted in terms of the mapping of permutations of n symbols to binary trees of height k. An explicit formula for this number is given with an equivalent recursive definition for computational use. A generating function is derived for the number of symbols given height. Lower and upper bounds on random binary tree height are developed and shown to approach one another asymptotically as a function of n, providing a limiting expression for the expected height.

The random binary trees are examined further to provide expressions for the expectations of the number of vacancies at each level, the distribution of vacancies over all levels, the comparisons required for insertion of a new random symbol, the fraction of nodes occupied at a particular level, the number of leaves, the number of single vacancies at each level, and the number of twin vacancies at each level. A random process is defined for the number of symbols required to grow a tree exceeding any given height.

Finally, an appendix is given with sample tabulations and figures of the distributions.

The distributional results yield a hypothesis test for randomness. Such tests are of particular interest in areas such as computer science and cryptoanalysis, where transformations which introduce nonrandomness must be efficiently screened to identify the transformation groups associated with such behavior.

The class of random trees which we investigate is quite distinct from the widely-studied *equally-likely* trees common in the literature. We have never encountered an application leading to this latter class of trees, and suspect that its popularity is founded at least in part on the relative ease of analysis.

The mathematical tools brought to bear on this problem (especially regarding the asymptotic behavior of sequences of functions) are not commonly used in the operations research literature, but are recommended to those interested in distributional as well as worst-case complexity arguments.

2. Number of binary trees of a given height. In this section we consider the problem of finding the number of binary trees with n nodes and height k. Denote this number by $t(n,k)$, where n and k are positive integers. Since $t(n,k) = 0$ unless

$$k \leqslant n \leqslant 2^k - 1 \tag{2.1}$$

we are only concerned with integers $n = 1, 2, \ldots$; $k = 1, 2, \ldots$ satisfying the inequality (2.1).

An explicit formula for the numbers $t(n,k)$ can be obtained by the following simple combinatorial argument. Consider the class of all binary trees with n nodes and height k which have exactly m_j nodes at the level $j+1$, $j = 1, \ldots, k-1$. Let $m_j = l_j + r_j$, where l_j and r_j are numbers of nodes which are left successors and right successors of nodes at the level j. In other words, l_j is the number of nodes at level $j+1$ at the end of left going arcs emanating from nodes at level j. These can be selected in $\binom{m_j}{l_j}$ ways, and the r_j nodes in $\binom{m_j}{r_j}$ ways. Thus, the total number of ways to arrange the arcs between nodes at levels j and $j+1$ is given by

$$\sum_{\substack{l_j \geqslant 0 \\ l_j + r_j = m_j}} \sum_{r_j \geqslant 0} \binom{m_{j-1}}{l_j} \binom{m_{j-1}}{r_j} = \binom{2m_{j-1}}{m_j}. \tag{2.2}$$

Since $m_0 = 1$ (the root) and $m_1 + \cdots + m_k = n - 1$ with $m_j \geqslant 1$ for $j = 1, \ldots, k$ we

obtain from (2.2) the formula

$$t(n,k) = \sum \binom{2}{m_1}\binom{2m_1}{m_2} \cdots \binom{2m_{k-2}}{m_{k-1}}, \tag{2.3}$$

where the summation is over all integers m_j, $j = 1, \ldots, k$ satisfying

$$m_j \geqslant 1, \qquad j = 1, \ldots, k-1, \quad \text{and} \quad m_1 + \cdots + m_{k-1} = n - 1.$$

The formula is valid for $n > 1$ and k satisfying (2.1), for $n = 1$ we have trivially $t(1,1) = 1$.

Although (2.3) is an explicit formula for the number $t(n,k)$ it is not very convenient for calculation. An alternate way is through a recurrence. Let

$$t(n,k) = T(n,k) - T(n,k-1), \qquad n \geqslant 1, \quad k \geqslant 1, \tag{2.4}$$

where $T(n,k)$ is the number of binary trees with n nodes and height not exceeding k. If we define

$$T(n,0) = \begin{cases} 1 & \text{if} \quad n = 0, \\ 0 & \text{if} \quad n > 0, \end{cases} \tag{2.5}$$

and $T(0,k) = 1$ for $k \geqslant 0$, we obtain the recurrence relation

$$T(n+1,k+1) = \sum_{j=0}^{n} T(j,k)T(n-j,k), \tag{2.6}$$

valid for $n \geqslant 0$ and $k \geqslant 0$. This follows from the fact that the class of all binary trees with $n+1$ nodes and height not exceeding $k+1$ can be partitioned into $n+1$ subclasses according to the number of nodes j in the left subtree of root. Since the heights of both the left and right subtrees must not exceed k the number of trees in the jth class is the product $T(j,k)T(n-j,k)$, and (2.6) follows.

Note that with the convention (2.5) the recurrence (2.6) yields automatically $T(n,k) = 0$ for $n > 2^k - 1$, and that for $0 \leqslant n \leqslant k$, $T(n,k)$ is just the number of binary trees with n nodes. It is well known (see Knuth [8]) that the latter are Catalan numbers

$$C_n = \frac{1}{n+1}\binom{2n}{n}, \qquad n \geqslant 0, \tag{2.7}$$

so that

$$T(n,k) = C_n \qquad \text{for} \quad 0 \leqslant n \leqslant k. \tag{2.8}$$

From the recurrence (2.6) one easily obtains the sequence enumerators defined by

$$f_k(x) = \sum_{n \geqslant 0} T(n,k)x^n, \qquad k \geqslant 0. \tag{2.9}$$

Since the right-hand side of (2.6) is the Cauchy product, we have immediately

$$\sum_{n \geqslant 0} T(n+1,k+1)x^n = f_k^2(x),$$

from which in view of (2.5) we obtain

$$\left.\begin{aligned} f_{k+1}(x) &= 1 + xf_k^2(x), \qquad k \geqslant 0, \\ &\text{with} \quad f_0(x) = 1. \end{aligned}\right\} \tag{2.10}$$

Note that $f_k(x)$ is a polynomial in x, hence if x is regarded as a complex variable f_k, $k = 0, 1, \ldots,$ is a sequence of entire functions.

We now show that this sequence converges uniformly in a circular region of the complex plane, specifically that as $k \to \infty$

$$f_k(z) \to \omega(z) \qquad \text{uniformly for} \quad |z| \leqslant \tfrac{1}{4}, \quad \text{where} \tag{2.11}$$

$$\omega(z) = \sum_{n=0}^{\infty} \frac{1}{n+1}\binom{2n}{n} z^n = \frac{1-\sqrt{1-4z}}{2z}. \tag{2.12}$$

To see this, note that with C_n as in (2.7) we have $T(n,k) \leqslant C_n$ for all k which together with (2.8) yields

$$\omega(z) - f_k(z) = z^k \sum_{n>k} (C_n - T(n,k)) z^{n-k},$$

so that for $|z| \leqslant \frac{1}{4}$

$$|\omega(z) - f_k(z)| \leqslant \sum_{n>k} C_n |z|^n \leqslant \sum_{n>k} C_n 4^{-n},$$

which is a tail of the expansion $\omega(\frac{1}{4}) = 2$.

This result will now be used to develop an asymptotic distribution of the numbers $t(n,k)$ as $k \to \infty$. In doing this, we follow the method of Renyi and Szekeres used in [9] for a similar problem.

From (2.4) we have for $k \geqslant 1$, $t(0,k) = 0$,

$$\sum_{n>0} t(n,k) z^n = f_k(z) - f_{k-1}(z), \tag{2.13}$$

which are entire functions for every $k \geqslant 1$. Hence by the Cauchy formula

$$t(n,k) = \frac{1}{2\pi i} \oint \frac{f_k(z) - f_{k-1}(z)}{z^{n+1}}\, dz, \tag{2.14}$$

where we take the circle $|z| = \frac{1}{4}$ as the contour of integration. To estimate the integral we use Laplace's method by first showing that as $k \to \infty$ the only significant contribution of the integrand is in the vicinity of positive real axis.

For $z = \frac{1}{4}e^{i\phi}$ the function $\omega(z)$ defined by (2.12) becomes

$$\omega\left(\tfrac{1}{4}e^{i\phi}\right) = 2e^{-i\phi}\left(1 - \sqrt{1 - e^{i\phi}}\right), \qquad -\pi \leqslant \phi \leqslant \pi. \tag{2.15}$$

Its modulus $2|1 - \sqrt{1 - e^{i\phi}}|$ is a decreasing function of $|\phi|$ on $0 \leqslant |\phi| \leqslant \pi$ with maximum 2 at $\phi = 0$ and minimum $2(\sqrt{2} - 1)$ at $|\phi| = \pi$.

Now use (2.10) to write

$$f_{k+1}(z) - \omega(z) = z\left[f_k(z) - \omega(z)\right]\left[f_k(z) + \omega(z)\right] \tag{2.16}$$

and choose an arbitrary $0 < \epsilon < 2/3$. By (2.11) there exists K_ϵ such that

$$k \geqslant K_\epsilon \Rightarrow |f_k(z) - \omega(z)| < \epsilon \qquad \text{for all } \phi \tag{2.17}$$

whence using (2.16)

$$\begin{aligned} |f_{k+1}(z) - \omega(z)| &< \epsilon\left(2|\omega\left(\tfrac{1}{4}e^{i\phi}\right)| + \epsilon\right) \\ &\leqslant 4\epsilon|1 - \sqrt{1 - e^{i\phi}}|\left(1 + \frac{\epsilon}{2|\omega\left(\frac{1}{4}e^{i\phi}\right)|}\right) \\ &< 12\epsilon|1 - \sqrt{1 - e^{i\phi}}| \quad \text{since} \quad |\omega\left(\tfrac{1}{4}e^{i\phi}\right)| > \tfrac{2}{3}. \end{aligned} \tag{2.18}$$

Iterating (2.16) we thus find that

$$|f_{K+m}(z) - \omega(z)| < \left(12\epsilon\left|1 - \sqrt{1 - e^{i\phi}}\right|\right)^m \tag{2.19}$$

for all $m = 1, 2, \ldots$ and all ϕ, i.e., that

$$f_k\left(\tfrac{1}{4}e^{i\phi}\right) = \omega\left(\tfrac{1}{4}e^{i\phi}\right) + o\left(\left|1 - \sqrt{1 - e^{i\phi}}\right|^k\right) \tag{2.20}$$

uniformly in ϕ.

Since for sufficiently small $|\phi|$ we have

$$\left|1 - \sqrt{1 - e^{i\phi}}\right| \leqslant 1 - \tfrac{1}{2}\sqrt{|\phi|} \tag{2.21}$$

we can choose for instance a sequence $\phi_k = (\ln^2 k/k)^2$ for which as $k \to \infty$

$$\sup_{|\phi| \geqslant \phi_k} \left|1 - \sqrt{1 - e^{i\phi}}\right|^k \leqslant e^{-\ln^2 k} \tag{2.22}$$

and thus conclude that for $|\phi| \geqslant \phi_k$

$$f_k\left(\tfrac{1}{4}e^{i\phi}\right) = \omega\left(\tfrac{1}{4}e^{i\phi}\right) + o(e^{-n^2 k}) \tag{2.23}$$

as $k \to \infty$.

If we now substitute (2.23) into (2.14) we have with $z = \frac{1}{4}e^{i\phi}$

$$\begin{aligned} t(n,k) &= \frac{4^n}{2\pi}\int_{-\pi}^{\pi}\left[f_k\left(\frac{1}{4}e^{i\phi}\right) - f_{k-1}\left(\frac{1}{4}e^{i\phi}\right)\right]e^{-in\phi}\,d\phi \\ &= \frac{4^n}{2\pi}\int_{|\phi|<(\ln^2 k/k)^2}\left[f_k\left(\frac{1}{4}e^{i\phi}\right) - f_{k-1}\left(\frac{1}{4}e^{i\phi}\right)\right]e^{in\phi}\,d\phi + 4^n o(e^{-\ln^2 k}), \end{aligned} \tag{2.24}$$

as $k \to \infty$.

To estimate the remaining integral we first set $\phi = 0$ and call

$$\alpha_k = f_k\left(\tfrac{1}{4}\right). \tag{2.25}$$

The recurrence (2.10) yields

$$\alpha_{k+1} = 1 + \tfrac{1}{4}\alpha_k^2, \qquad \alpha_0 = 1 \tag{2.26}$$

from which by substitution $\gamma_k = 2 - \alpha_k$ we obtain the recurrence

$$\gamma_{k+1} = \gamma_k - \tfrac{1}{4}\gamma_k^2, \qquad \gamma_0 = 1. \tag{2.27}$$

Thus γ_k is the kth iterate of the function $h(x) = x - \frac{1}{4}x^2$, and iterates of such functions can be handled by standard methods (see, e.g., de Bruijn [4, §8.7 or Exercise 8.11]). We obtain the asymptotic expansion

$$\begin{aligned} \gamma_k &= \frac{4}{k} - \frac{4\ln k}{k^2} + \frac{c}{k^2} + O\left(\frac{\ln^2 k}{k^3}\right) \quad \text{or} \\ f_k\left(\frac{1}{4}\right) &= 2 - \frac{4}{k} + \frac{4\ln k}{k^2} - \frac{c}{k^2} + O\left(\frac{\ln^2 k}{k^3}\right) \end{aligned} \tag{2.28}$$

as $k \to \infty$ where c is a constant.

To obtain the expansion for $|\phi| < (\ln^2 k/k)^2$ we use the method of variable coeffi-

cients on (2.28). See Szekeres [12]. Put

$$f_k\left(\tfrac{1}{4}e^{i\phi}\right) = 2 - \frac{1}{k}\, g_1(\xi_k) + \frac{\ln k}{k^2}\, g_2(\xi_k) - \frac{c}{k^2}\, g_3(\xi_k) + \cdots \text{ (higher order terms)}. \tag{2.29}$$

where $\xi_k^2 = i\phi k^2$.

For ϕ fixed we then have $\xi_{k+1} = \xi_k + \xi_k/k$ so that upon substitution and using the expansion

$$g_1\left(\xi + \frac{1}{k}\xi\right) = g_1(\xi) + \frac{\xi}{k}\, g_1'(\xi) + \cdots,$$

$$g_2\left(\xi + \frac{1}{k}\xi\right) = g_2(\xi) + \frac{\xi}{k}\, g_2''(\xi) + \cdots,$$

$$\frac{1}{k+1} = \frac{1}{k}\,\frac{1}{1+k^{-1}} = \frac{1}{k} - \frac{1}{k^2} + \frac{1}{k^3} - \cdots,$$

$$\frac{1}{(k+1)^2} = \frac{1}{k^2}\,\frac{1}{(1+k^{-1})^2} = \frac{1}{k^2} - \frac{2}{k^3} + \cdots, \quad \text{and} \quad \ln(k+1) = \ln k + \frac{1}{k} - \cdots,$$

we obtain after collecting terms

$$f_{k+1}\left(\tfrac{1}{4}e^{i\phi}\right) = 2 - \frac{1}{k}\, g_1(\xi) \frac{\ln k}{k^2}\, g_2(\xi) + \frac{1}{k^2}\left[\, g_1(\xi) - \xi g_1'(\xi) - c g_3(\xi)\right]$$

$$- \frac{\ln k}{k^3}\left[2g_2(\xi) - g_2'(\xi)\right] + \cdots. \tag{2.30}$$

On the other hand from (2.10) $f_{k+1}(\frac{1}{4}e^{i\phi}) = 1 + \frac{1}{4}e^{i\phi} f_k^2(\frac{1}{4}e^{i\phi})$, whence using (2.29) and $e^{i\phi} = e^{(\xi/k)^2} = 1 + (\xi/k)^2 + \cdots$, we have

$$f_{k+1}\left(\tfrac{1}{4}e^{i\phi}\right) = 2 - \frac{1}{k}\, g_1(\xi) + \frac{\ln k}{k^2}\, g_2(\xi) + \frac{1}{k^2}\left[\, \frac{1}{4}\, g_1^2(\xi) + \xi^2 - c g_3(\xi)\right]$$

$$- \frac{\ln k}{k^3}\, \frac{1}{2}\, g_1(\xi)\, g_2(\xi) + \cdots. \tag{2.31}$$

Comparing terms of (2.30) and (2.31) we obtain

$$g_1(\xi) - \xi g_1'(\xi) = \tfrac{1}{4}\, g_1^2(\xi) + \xi^2, \qquad 2g_2(\xi) - \xi g_2'(\xi) = \tfrac{1}{2}\, g_1(\xi)\, g_2(\xi),$$

with initial conditions $g_1(0) = g_2(0) = 4$ from (2.28). These equations have a solution

$$g_1(\xi) = 2\xi \cot \tfrac{1}{2}\xi, \qquad g_2(\xi) = \xi^2 (\sin \tfrac{1}{2}\xi)^{-2},$$

which, when substituted back into (2.29), gives

$$f_k\left(\tfrac{1}{4}e^{i\phi}\right) = 2 - \frac{2}{k}\,\xi \cot \tfrac{1}{2}\xi + \frac{\ln k}{k^2}\,\xi^2 (\sin \tfrac{1}{2}\xi)^{-2} + O(k^{-2}),$$

$$\text{where} \quad \xi^2 = i\phi k^2, \quad |\phi| < \left(\frac{\ln^2 k}{k}\right)^2. \tag{2.32}$$

Next

$$f_k\left(\tfrac{1}{4}e^{i\phi}\right) - f_{k-1}\left(\tfrac{1}{4}e^{i\phi}\right) = \frac{2}{k-1}\,\xi_{k-1} \cot \tfrac{1}{2}\xi_{k-1} - \frac{2}{k}\,\xi_k \cot \tfrac{1}{2}\xi_k + \frac{\ln k}{k^2}\,\xi_k^2 (\sin \tfrac{1}{2}\xi_k)^{-2}$$

$$- \frac{\ln(k-1)}{(k-1)^2}\,\xi_{k-1}^2 (\sin \tfrac{1}{2}\xi_{k-1})^{-2} + O(k^{-3}),$$

and setting $\xi_{k-1} = \xi - \xi/k$, $\xi = \xi_k$, so that $\xi_{k-1}/(k-1) = \xi/k$, and using the expansions

$$\cot\tfrac{1}{2}\left(\xi - \frac{1}{k}\xi\right) = \cot\tfrac{1}{2}\xi + \frac{1}{2k}\xi\left(\sin\tfrac{1}{2}\xi\right)^{-2}$$

$$= \frac{\xi^2}{4k^2}\cot\tfrac{1}{2}\xi\left(\sin\tfrac{1}{2}\xi\right)^2 + \cdots,$$

$$\ln(k-1) = \ln k - \frac{1}{k} + \cdots, \quad \text{and}$$

$$\left(\sin\tfrac{1}{2}\left(\xi - \frac{1}{k}\xi\right)\right)^{-2} = \left(\sin\tfrac{1}{2}\xi\right)^{-2}\left(1 + \frac{1}{k}\xi\cot\tfrac{1}{2}\xi + \cdots\right),$$

we obtain

$$f_k\left(\tfrac{1}{4}e^{i\phi}\right) - f_{k-1}\left(\tfrac{1}{4}e^{i\phi}\right) = \frac{1}{k^2}\xi^2\left(\sin\tfrac{1}{2}\xi\right)^{-2} - \frac{\ln k}{k^3}\xi^3\cot\tfrac{1}{2}\xi\left(\sin\tfrac{1}{2}\xi\right)^{-2}$$

$$+ \frac{\xi^2}{k^3}\left(\sin\tfrac{1}{2}\xi\right)^{-2} - \frac{\xi^3}{4k^3}\cot\tfrac{1}{2}\xi\left(\sin\tfrac{1}{2}\xi\right)^{-2} + O(k^{-3}).$$

However, for $\xi^2 = i\phi k^2$, $|\phi| < (\ln^2 k/k)^2$, we have

$$\xi\cot\tfrac{1}{2}\xi = O(\ln^2 k), \qquad |\sin\tfrac{1}{2}\xi|^{-2} \to 0 \qquad \text{as} \quad k \to \infty,$$

so that we can write

$$f_k\left(\tfrac{1}{4}e^{i\phi}\right) - f_{k-1}\left(\tfrac{1}{4}e^{i\phi}\right) = \frac{1}{k^2}\,\frac{\xi^2}{\sin^2\tfrac{1}{2}\xi} + O\left(\frac{\ln^4 k}{k^3}\right). \tag{2.33}$$

The same expression can also be obtained by again substituting (2.32) into the recurrence (2.10). If we now insert (2.33) into the integral in (2.24) and use the substitution $\xi^2 = i\phi k^2$, (2.24) becomes

$$t(n,k) = \frac{4^n}{2\pi i}\,\frac{2}{k^4}\int_\Gamma \frac{\xi_3}{\sin^2\tfrac{1}{2}\xi}\,e^{-n\xi^2/k^2}d\xi + 4^n O\left(\frac{\ln^8 k}{k^5}\right), \tag{2.34}$$

where the path of integration Γ as determined from the condition $|\phi| < (\ln^2 k/k)^2$ is (see Figure 2.2):

$$\xi = \tau(-1+i) \qquad \text{for} \quad -\frac{1}{\sqrt{2}}\ln^2 k < \tau \leqslant 0,$$

$$\xi = \tau(1+i) \qquad \text{for} \quad 0 \leqslant \tau < \frac{1}{\sqrt{2}}\ln^2 k.$$

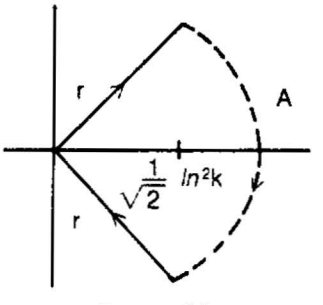

FIGURE 2.2

The integrand is a meromorphic function with poles at

$$\xi_m = 2\pi m, \qquad m = \pm 1, \pm 2, \ldots, \tag{2.35}$$

and corresponding residues

$$r(\xi_m) = 4\left[3(2\pi m)^2 - 2\frac{n}{k^2}(2\pi m)^4\right]e^{-n(2\pi m)^2/k^2}. \tag{2.36}$$

Before applying the residue theorem, we have to close the path of integration, for instance by the arc (see Figure 2.2) A: $\xi = \rho_k e^{-i\theta}$, $\rho_k = \ln^2 k$, $-\pi/4 \leqslant \theta \leqslant \pi/4$. Deforming the arc in the neighborhood of the real axis so that it passes approximately in between the two consecutive poles we have from $|\sin z|^2 = \sin^2 x + \sinh^2 y$

$$|\sin \tfrac{1}{2}\xi|^{-2} \leqslant 1 + \eta, \qquad \eta > 0,$$

along the arc for large enough k. Hence

$$\begin{aligned}\left|\int_A \frac{\xi^3}{\sin^2 \frac{1}{2}\xi} e^{-n\xi^2/k^2}\, d\xi\right| &\leqslant (1+\eta)\rho_k^4 \int_0^{\pi/2} e^{-(n/k^2)(\rho_k^2 \sin\theta)}\, d\theta \\ &\leqslant (1+\eta)\rho_k^4 \int_0^{\pi/2} e^{-(2n/k^2)(\rho_k^2\theta)}\, d\theta \\ &= \frac{(1+\eta)\pi\rho_k^2 k^2}{2n}\left(1 - e^{-n\rho_k^2/k^2}\right). \\ &\leqslant \frac{(1+\eta)\pi\rho_k^2 k^2}{2n}\,\frac{n}{k^2}\rho_k^2 = O(\ln^8 k) \qquad \text{as} \quad k \to \infty. \end{aligned} \tag{2.37}$$

Finally, applying the residue theorem to the integral over the contour $\Gamma + A$ and using (2.34), (2.36) and (2.37) we have the asymptotic expansion

$$\begin{aligned} t(n,k) = 4^{n+1}\frac{2}{k^4}\sum_{m\geqslant 1}\left[2\frac{n}{k^2}(2\pi m)^4 - 3(2\pi m)^2\right]e^{-n(2\pi m)^2/k^2} \\ + 4^n O(e^{-\ln^2 k}) + 4^n O\left(\frac{\ln^8 k}{k^5}\right) + 4^n O\left(\frac{\ln^8 k}{k^4}\right). \end{aligned} \tag{2.38}$$

From here we can get the distribution of heights of binary trees with n nodes or equivalently the probability that a randomly selected binary tree has height k if all binary trees with n nodes are considered equally likely.

Calling this quantity $p_n(k)$ we have

$$p_n(k) = \frac{t(n,k)}{C_n}, \tag{2.39}$$

where $C_n = \sum_k t(n,k)$ is the Catalan number (2.7). Using the asymptotic expansion [7] $C_n = 4^n/n^{3/2}\sqrt{\pi} + O(4^n n^{-5/2})$ we obtain for large k and n,

$$\begin{aligned} P_n(k) = 8\sqrt{\frac{\pi}{n}}\,\beta^2 \sum_{m\geqslant 1}\left[2\beta(2\pi m)^4 - 3(2\pi m)^2\right]e^{-(2\pi m)^2\beta} \\ + O(n^{3/2}e^{-\ln^2 k}) + O\left(\frac{n^{3/2}\ln^8 k}{k^5}\right) + O\left(\frac{n^{3/2}\ln^8 k}{k^4}\right), \end{aligned} \tag{2.40}$$

where $\beta = n/k^2$. Note that if $k \leqslant n$, which is the case of interest, the dominant O-term

is the last one. Thus, if for some arbitrarily small $\delta > 0$

$$n^{3/8+\delta} \leqslant k \leqslant n, \tag{2.41}$$

then as $n \to \infty$

$$P_n(k) \sim 8\sqrt{\frac{\pi}{n}}\ \beta^2 \sum_{m=1}^{\infty} \left[2\beta(2\pi m)^4 - 3(2\pi m)^2\right] e^{-(2\pi m)^2\beta}. \tag{2.42}$$

To verify that (2.42) is indeed a probability distribution note that as $n \to \infty$

$$\sum_k p_n(k) \sim 8\sqrt{\frac{\pi}{n}} \int \beta^2 \sum_{m=1}^{\infty} \left[2\beta(2\pi m)^4 - 3(2\pi m)^2\right] e^{-(2\pi m)^2\beta}\, dk,$$

where the limits of summation and integration are given by (2.41). Making a substitution $k = \sqrt{n/\beta}$ we have as $n \to \infty$

$$4\sqrt{\pi} \sum_{m=1}^{\infty} \left\{ (2\pi m)^4 \int_0^{\infty} 2\beta^{3/2} e^{-(2\pi m)^2\beta}\, d\beta - (2\pi m)^2 \int_0^{\infty} 3\beta^{1/2} e^{-(2\pi m)^2\beta}\, d\beta \right\}.$$

But

$$\int_0^{\infty} 3\beta^{1/2} e^{-(2\pi m)^2\beta}\, d\beta = 2\beta^{3/2} e^{-(2\pi m)^2\beta}\Big|_0^{\infty} + (2\pi m)^2 \int_0^{\infty} 2\beta^{3/2} e^{-(2\pi m)^2\beta}\, d\beta,$$

and we are left with

$$\lim_{\beta \to 0+} 4\sqrt{\pi} \sum_{m=1}^{\infty} 2(2\pi m)^2 \beta^{3/2} e^{-(2\pi m)^2\beta} = \frac{4}{\sqrt{\pi}} \int_0^{\infty} \mu^2 e^{-\mu^2}\, d\mu = 1$$

by the substitution $\mu^2 = (2\pi m)^2\beta$. If we denote $\mathfrak{H}_n$ as the random variable with distribution $p_n(k)$ and call the asymptotic distribution function

$$F(x) = \lim_{n \to \infty} P\left(\frac{\mathfrak{H}_n}{2\sqrt{n}} \leqslant x \right) \tag{2.43}$$

we obtain by integrating the right-hand side of (2.42) over $0 \leqslant k \leqslant 2x\sqrt{n}$ by the same method as above

$$F(x) = 4x^{-3}\pi^{5/2} \sum_{m=1}^{\infty} m^2 e^{-(\pi m/x)^2}. \tag{2.44}$$

It can easily be seen from (2.43) that for $x = O(\sqrt{n})$, $F(x) \to 0$ and thus the asymptotic distribution (2.42) or equivalently (2.43), (2.44) is valid for all k in the range (2.1) as $n \to \infty$ (see Appendix, Figure A.1). This is identical with the distribution function obtained by Renyi and Szekeres (see [9, p. 506]) for the height of general (as opposed to binary) trees. Only the normalizing factor in (3.43) differs by a constant, namely $\sqrt{2}$. More precisely, if $\mathfrak{H}_n^*$ is the height of a general tree with n nodes (i.e., with no restriction on the number of successors of a node) then for large n

$$\mathfrak{H}_n \sim \sqrt{2}\, \mathfrak{H}_n^*, \tag{2.45}$$

a somewhat surprising result.

3. Binary trees generated by random permutations. As described in the introduction the tree insertion algorithm defines a map which assigns to every permutation $\pi(1, \ldots, n)$ a binary tree with n nodes. If all permutations of integers $1, \ldots, n$ are

considered equally likely the resulting trees are referred to as random. Thus, every numerical quantity defined on a binary tree becomes a random variable.

Let $\mathscr{H}_n$ be the height of a random binary tree with n nodes, i.e., generated by random permutations of the first n integers. Clearly, the probability

$$P(\mathscr{H}_n \leqslant k) = \frac{1}{n!} B(n,k), \tag{3.1}$$

where $B(n,k)$ is the number of permutations $\pi(1, \ldots, n)$ mapped into trees with height not exceeding k (see Appendix, Table A.2).

In order to obtain an explicit expression for (3.1) we first need a suitable indexing system for the nodes. A natural way to do this is to consider first a full binary tree and label its nodes by the sequence of positive integers starting from the root and labelling in each subsequent level from left to right (see Figure. 3.1).

Thus nodes at level j, $j = 1, 2, \ldots,$ from left to right have labels

$$2^{j-1} + m, \qquad m = 0, \ldots, 2^{j-1}. \tag{3.2}$$

Note that the left and right successors of a node labelled x have labels $2x$ and $2x + 1$, respectively, left successors always have even labels, right successors have odd labels (the root being an exception).

Next consider a fixed binary tree with n nodes and for every label x of the form (3.2) define

$$d(x) = \begin{cases} 0 & \text{if there is no node with label } x, \\ 1 + d(2x) + d(2x+1) & \text{if there is a node with label } x. \end{cases} \tag{3.3}$$

Note that $d(x)$ is simply the number of nodes in a subtree with root at x, in particular $d(1) = n$ and $d(x) = 1$ if and only if the node labelled x has no successor, i.e., is a leaf.

LEMMA. *There is a one-to-one correspondence between binary trees with n nodes and height not exceeding k and the set $D(n,k)$ of vectors $(d(1), d(2), \ldots, d(2^k - 1))$ with nonnegative integral components satisfying the conditions*:

(1) $d(1) = n$,

(2) *for all* $j = 1, \ldots, k$

$$2^{j-1} \leqslant x \leqslant 2^j - 1 \Rightarrow 0 \leqslant d(x) \leqslant 2^{k-j+1} - 1,$$

(3) $d(2x) + d(2x+1) > 0 \Rightarrow d(x) = 1 + d(2x) + d(2x+1)$.

PROOF. Given a binary tree the numbers $d(x)$ are uniquely defined by (3.3). Property (1) is obvious, property (3) follows from (3.3) since nodes $2x$ and $2x + 1$ are successors of node x, hence if there is a node with labels either $2x$ or $2x + 1$ there must be one with label x. Property (2) is necessary since if it were violated then there would be a subtree with root at level j having at least 2^{k-j+1} nodes and the height of the tree would then exceed k. Conversely, given a vector $(d(1), \ldots, d(2^k - 1))$ satisfying (1)–(3), construct first a complete binary tree of height k and then eliminate all nodes with labels x such that $d(x) = 0$. Properties (1)–(3) then guarantee that the result is a binary tree with exactly n nodes. ∎

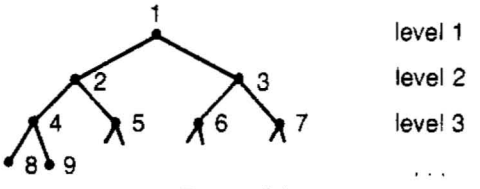

FIGURE 3.1

A vector $(d(1), \ldots, d(2^k - 1)) \in D(n,k)$ can be used to compute the number of permutations mapped into the tree corresponding to this vector. Consider a permutation $\pi(1, \ldots, n) = (s_1, \ldots, s_n)$ mapped into a binary tree with height not exceeding k and look at the node into which a particular symbol, say s_i, is mapped. Let x_i be the label of this node and let $L(s_i)$ and $R(s_i)$ be the ordered subsets of $(s_1, \ldots, s_n)$, which are mapped into the left and right subtrees of the node x_i. By the very nature of the mapping if the elements of $L(s_i) \cup R(s_i)$ are reshuffled the tree is not changed as long as the order of elements within each subset $L(s_i)$ and $R(s_i)$ of $L(s_i) \cup R(s_i)$ is preserved. But the numbers of elements in $L(s_i)$ and $R(s_i)$ are $d(2x_i)$ and $d(2x_i + 1)$ respectively so that the number of permutations resulting from such a reshuffle is

$$\frac{\left[d(2x_i) + d(2x_i + 1)\right]!}{d(2x_i)!d(2x_i + 1)!} = \frac{\left[d(x_i) - 1\right]!}{d(2x_i)!d(2x_i + 1)!},$$

using (3.3) and $d(x_i) > 0$ (since s_i is mapped into a node labelled x_i). Repeating this argument for each node of the tree corresponding to $(d(1), \ldots, d(2^k - 1))$ gives the total number of permutations

$$\prod \frac{\left[d(x) - 1\right]!}{d(2x)!d(2x + 1)!},$$

where the product is over all $d(x) > 0$, $x = 1, \ldots, 2^{k-1} - 1$. From here using the lemma and (3.1) we obtain the formula

$$P(\mathscr{H}'_n \leqslant k) = \frac{1}{n!} \sum_{\mathbf{d} \in D(n,k)} \prod \frac{\left[d(x) - 1\right]!}{d(2x)!d(2x + 1)!}. \tag{3.4}$$

Note that if $\mathbf{d}' = (d'(1), \ldots, d'(2^k - 1))$ is obtained from $\mathbf{d} = (d(1), \ldots, d(2^k - 1))$ by

$$d'(x) = \begin{cases} d(x) & \text{if } d(x) > 0, \\ 1 & \text{if } d(x) = 0, \end{cases} \tag{3.5}$$

the formula (3.4) takes on a simpler form

$$P(\mathscr{H}'_n \leqslant k) = \sum_{\mathbf{d} \in D(n,k)} \left(\prod_{x=1}^{2^k - 1} d'(x) \right)^{-1}. \tag{3.6}$$

Unfortunately, except in a few special cases, the set $D(n,k)$ is quite complicated for (3.6) to be useful for computation. Again a recurrence relation may be preferable.

Indeed, such a recurrence is quite easy to derive. Consider the left and right subtrees of the root of a random binary tree with n nodes. If J is a random number of nodes in the left subtree then we have

$$\mathscr{H}'_{n+1} = 1 + \max\{\mathscr{H}'_J, \mathscr{H}'_{n-J}\}. \tag{3.7}$$

Now J is simply the number of symbols in a random permutation $\pi(1, \ldots, n + 1) = (s_1, \ldots, s_{n+1})$, which are less than s_1. Therefore, J is uniformly distributed over $\{0, \ldots, n\}$, and $\mathscr{H}'_J$ and $\mathscr{H}'_{n-J}$ are conditionally independent given J. Consequently

$$P(\mathscr{H}'_{n+1} - 1 \leqslant k \mid J = j) = P(\mathscr{H}'_j \leqslant k) P(\mathscr{H}'_{n-j} \leqslant k),$$

from which by calling

$$F(n,k) = P(\mathscr{H}'_n \leqslant k) \tag{3.8}$$

and taking the expectation we obtain the recurrence

$$\left.\begin{aligned} F(n+1,k+1) &= \frac{1}{n+1}\sum_{j=0}^{n} F(j,k)F(n-j,k) \\ &\text{valid for} \quad n \geqslant 0, \quad k \geqslant 0 \\ \text{if we define} \quad F(0,k) &= \begin{cases} 1 & \text{if} \quad k = 0, \\ 0 & \text{if} \quad k > 0. \end{cases} \end{aligned}\right\} \tag{3.9}$$

It may be noted that (3.9) gives

$$\left.\begin{aligned} F(n,k) &= 1 \qquad \text{for} \quad 0 \leqslant n \leqslant k, \quad \text{and} \\ F(n,k) &= 0 \qquad \text{for} \quad n \geqslant 2^k, \quad \text{as expected.} \end{aligned}\right\} \tag{3.10}$$

Now (3.9) can be used to calculate the numbers $F(n,k)$ for moderate values of k and n (see Appendix, Tables A.2 and A.3, and [2]); however the memory requirements increase rapidly.

It can also be used together with (3.10) for some special choice of n and k. For instance for n close to $2^k - 1$ we obtain

$$F(2^k - 1, k) = \prod_{j=1}^{k} \left(\frac{1}{2^j - 1}\right)^{2^{k-j}}, \qquad k \geqslant 1, \tag{3.11a}$$

$$F(2^k - 2, k) = (2^k - 1)F(2^k - 1, k), \qquad k \geqslant 1, \tag{3.11b}$$

$$F(2^k - 3, k) = (2^k - 1)(2^{k-1} - 1)F(2^k - 1, k), \qquad k \geqslant 2, \tag{3.11c}$$

and a few more. Note that (3.11a) is the probability of obtaining a full tree of height k. However, these few terms are of minor interest.

Another possible approach is to use generating functions. Defining

$$f_k(x) = \sum_{n=0}^{\infty} F(k,n)x^n, \qquad k \geqslant 0, \tag{3.12}$$

we have immediately from (3.9) the relation

$$f_{k+1}(x) = 1 + \int_0^x f_k^2(y)\,dy, \qquad f_0(x) = 1. \tag{3.13}$$

Note that $f_k(x)$ are again polynomials, $f_k(0) = 1$, and that for all $x \in (0, 1]$

$$f_k(x) < f_{k+1}(x) \to (1-x)^{-1} \qquad \text{as} \quad k \to \infty.$$

It is hoped that (3.13) can be used to obtain an asymptotic distribution of the heights $\mathscr{H}_n$. However, we have not been successful in that respect to date.

It has been suggested to us by A. Washburn that a lower bound on the expected height $E\{\mathscr{H}_n\}$ can be easily obtained from (3.7). Taking expectation we have

$$E\{\mathscr{H}_{n+1}\} = 1 + E\{\max(\mathscr{H}_J, \mathscr{H}_{n-J})\} \tag{3.14}$$

and conditioning upon J

$$E\{\max(\mathscr{H}_J, \mathscr{H}_{n-J}) \mid J = j\} \geqslant \max(E\{\mathscr{H}_j\}, E\{\mathscr{H}_{n-j}\}).$$

Since J is uniformly distributed over $\{0, \ldots, n\}$ this implies

$$E\{\max(\mathscr{H}_J, \mathscr{H}_{n-J})\} \geqslant \frac{1}{n+1}\sum_{j=0}^{n} \max(E\{\mathscr{H}_j\}, E\{\mathscr{H}_{n-j}\}). \tag{3.15}$$

Hence if α_n, $n = 0, 1, \ldots,$ is a sequence of numbers defined recursively by

$$\alpha_{n+1} = 1 + \frac{1}{n+1} \sum_{j=0}^{n} \max(\alpha_j, \alpha_{n-j}), \qquad \alpha_0 = 0, \tag{3.16}$$

it is easily seen from (3.14) and (3.15) that

$$E\{\mathscr{H}_n\} \geqslant \alpha_n \qquad \text{for all} \quad n = 0, 1, \ldots . \tag{3.17}$$

Note that α_n is a strictly increasing sequence so that

$$\sum_{j=0}^{n} \max(\alpha_j, \alpha_{n-j}) = \begin{cases} 2 \sum\limits_{n/2 < j \leqslant n} \alpha_j & \text{for } n \text{ odd}, \\ \alpha_{n/2} + 2 \sum\limits_{n/2 < j \leqslant n} \alpha_j & \text{for } n \text{ even}. \end{cases}$$

Since clearly $\alpha_n = O(n)$ we have as $n \to \infty$

$$\alpha_n \sim 1 + \frac{2}{n} \sum_{n/2 < j \leqslant n} \alpha_j,$$

which upon approximating the sum by an integral yields

$$\alpha(t) \sim 1 + 2\int_{1/2}^{1} \alpha(ty)\, dy. \tag{3.18}$$

On the other hand, let Y be a random variable defined by

$$nY = \begin{cases} J & \text{if} \quad \mathscr{H}_J \geqslant \mathscr{H}_{n-J}, \\ n - J & \text{if} \quad \mathscr{H}_J < \mathscr{H}_{n-J}. \end{cases}$$

Then

$$nY \leqslant \max\{J, n - J\}, \tag{3.19}$$

and by (3.7)

$$E\{\mathscr{H}_{n+1}\} = 1 + E\{\mathscr{H}_{nY}\}. \tag{3.20}$$

Now for large n, Y is uniformly distributed over $(\frac{1}{2}, 1)$, and since $\mu(n) = E\{\mathscr{H}_n\}$ is an increasing function of n we get from (3.20) by conditioning on Y and applying (3.19) the asymptotic inequality

$$\mu(n) \leqslant 1 + 2\int_{1/2}^{1} \mu(ny)\, dy \tag{3.21}$$

valid as $n \to \infty$. But this together with (3.17) and (3.18) indicates that we should have in fact

$$\mu(n) \sim \alpha(n) \qquad \text{as} \quad n \to \infty \tag{3.22}$$

where $\alpha(t)$ is a solution of (3.18). It is easily verified that (3.18) has a solution

$$\alpha(t) = \frac{\ln t}{1 - \ln 2} + \sigma,$$

where σ is an arbitrary constant. But as $\mu(n) \to \infty$ the constant can simply be disregarded and we have an asymptotic equivalence

$$\mu(n) \sim (1 - \ln 2)^{-1} \ln n = 3.25889 \ln n \tag{3.23}$$

as $n \to \infty$ (see Appendix, Tables A.2 and A.3).

Since completion of this work, Robson [10] has shown that asymptotically $3.63 \leqslant \mu(n)/\ln n \leqslant 4.31$.

4. Further results for random binary trees. We now turn our attention to some other quantities associated with random binary trees. In particular we wish to study the process of growing the tree as the number of nodes increases. A convenient way to visualize such a process is to consider a sequence $X_1, X_2, \ldots$ of independent random variables with a common continuous distribution and take as our permutation the ranks of $X_1, \ldots, X_n$ for each $n = 1, 2, \ldots$. Clearly, for each fixed n all resulting permutations are equally likely, and each new symbol, i.e., rank of X_{n+1}, is uniformly distributed over $\{1, \ldots, n+1\}$.

If for some n we have a random binary tree then a new symbol will be mapped into a new node. This new node can appear in one of $n+1$ possible locations in the current tree. Following Knuth [7] we will refer to these locations as vacancies of the current tree. They are depicted in Figure 4.1 as empty squares.

Let for $n \geqslant 1$, $j \geqslant 2$, $V_{n,j}$ be the number of vacancies at level j at time n, i.e., in a random binary tree with n nodes. Clearly $V_{n,j} = 0$ for $j > n+1$ and

$$V_{n,2} + V_{n,3} + \cdots + V_{n,n+1} = n + 1. \tag{4.1}$$

Also, obviously $V_{1,2} = 2$. Suppose now that at time n the vacancies have values $v_{n,2}, \ldots, v_{n,n+1}$. Since a new node is equally likely to fill any of the current $n+1$ vacancies, the probability that it will fill a vacancy at level j equals $v_{n,j}/(n+1)$. But then

$$\begin{aligned} V_{n+1,j} &= v_{n,j} - 1, \\ V_{n+1,j+1} &= v_{n,j+1} + 2, \qquad \text{while} \\ V_{n+1,i} &= v_{n,i} \qquad\qquad \text{for all } i \neq j,\ i \neq j+1. \end{aligned}$$

Thus if we define a sequence of random polynomials

$$W_n(x) = \sum_{j \geqslant 2} V_{n,j} x^{j-1}, \qquad n \geqslant 1,$$

we see that these polynomials form a Markov chain with transition probabilities

$$P\big(W_{n+1}(x) = w_n(x) + x^{j-1}(2x-1) \,|\, W_n(x) = w_n(x)\big) = \frac{v_{n,j}}{n+1},$$

$$n \geqslant 1, \quad j = 2, \ldots, n+1, \quad \text{and the initial state} \quad W_1(x) = 2x.$$

In particular we then have

$$\begin{aligned} E\{W_{n+1}(x) \,|\, W_n(x)\} &= W_n(x) + \sum_{j=2}^{n+2} x^{j-1}(2x-1)\frac{v_{n,j}}{n+1} \\ &= W_n(x)\left[1 + \frac{2x-1}{n+1}\right], \qquad n \geqslant 1. \end{aligned}$$

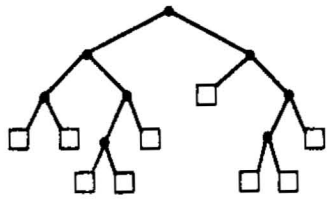

FIGURE 4.1

Consequently

$$E\{W_n(x)\} = \left(1 + \frac{2x-1}{n}\right)\left(1 + \frac{2x-1}{n-1}\right)\cdots\left(1 + \frac{2x-1}{2}\right)2x$$

$$= \frac{1}{n!}\,2x(2x+1)\cdots(2x+n-1) = \frac{1}{n!}\sum_{j=2}^{n+1}\begin{bmatrix} n \\ j-1 \end{bmatrix}(2x)^{j-1}, \quad (4.2)$$

where $[{}^{n}_{m}]$ are Stirling numbers of the first kind in Knuth's definition and notation [7]. Thus

$$E\{V_{nj}\} = \frac{2^{j-1}}{n!}\begin{bmatrix} n \\ j-1 \end{bmatrix}, \qquad j = 2, \ldots, n+1, \quad n \geqslant 1, \quad (4.3)$$

gives us the expectation of the number of vacancies at various levels (see Appendix, Table A.5). Other quantities of some interest may be obtained from here. For instance the expected distribution of the number of vacancies over levels is immediate

$$v_n(j) = \frac{1}{n+1}E\{V_{nj}\} = \frac{2^{j-1}}{(n+1)!}\begin{bmatrix} n \\ j-1 \end{bmatrix}, \qquad j = 2, \ldots, n+1; \quad n \geqslant 1. \quad (4.4)$$

Next let K_n be the number of comparisons needed to insert a new $(n+1)$st symbol into a binary tree by the algorithm described in the introduction. Then the probability

$$P(K_n = \kappa) = E\{P(K_n = \kappa \mid W_n(x))\}$$

$$= \left\{\frac{V_{n,\kappa+1}}{n+1}\right\} = v_n(\kappa+1), \qquad \kappa = 1, \ldots, n, \quad n \geqslant 1,$$

since the numbers of comparisons is κ if and only if the symbol fills a vacancy at level $\kappa+1$. Recalling that the Stirling number $[{}^{n}_{\kappa}]$ is also the number of permutations of n symbols with exactly κ cycles we can write

$$P(K_n = \kappa) = \frac{2^{\kappa}}{n+1}\,q_n(\kappa),$$

where $q_n(\kappa)$ is the probability that a random permutation of n symbols has κ cycles. (One wonders whether the number of comparisons can be related to the number of cycles.) (See Appendix, Table A.7.)

With

$$P(K_n = \kappa) = \frac{2^{\kappa}}{(n+1)!}\begin{bmatrix} n \\ \kappa \end{bmatrix}, \qquad \kappa = 1, \ldots, n,$$

the generating function (by definiton of the Stirling number) is

$$G_n(x) = \sum_{\kappa=1}^{n} P(k_n = \kappa)x^{\kappa} = \prod_{\kappa=1}^{n}\frac{2x+\kappa-1}{\kappa+1}.$$

Hence k_n is a convolution, $k_n = x_1 + \cdots + x_n$, where the x_κ's are independent Bernoulli variates:

$$x_\kappa = \begin{cases} 1, & p_\kappa = \dfrac{2}{\kappa+1}, \\ 0, & q_\kappa = \dfrac{\kappa-1}{\kappa+1}. \end{cases}$$

Consequently

$$\mu_n = E\{k_n\} = \sum_{\kappa=1}^{n} p_\kappa = 2\left(\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n+1}\right)$$
$$= 2(H_{n+1} - 1),$$

with H_n representing the Harmonic number [7]. (See also the Riemann Zeta function in Abramowitz and Stegun [1].)

$$\sigma_n^2 = \mathrm{Var}\{k_n\} = \sum_{\kappa=1}^{n} p_\kappa q_\kappa = 2\sum_{\kappa=1}^{n} \frac{1}{\kappa+1} - 4\sum_{\kappa=1}^{n} \frac{1}{(\kappa+1)^2}$$
$$= 2(H_{n+1} - 1) - 4\left(H_{n+1}^{(2)} - 1\right) = 2H_{n+1} - 4H_{n+1}^{(2)} + 2.$$

Since asymptotically [6]

$$H_n = \gamma + \ln n + \frac{1}{2n} + O\left(\frac{1}{n^2}\right), \qquad H_n^{(2)} = \frac{\pi^2}{6} - \frac{1}{n+1} + O\left(\frac{1}{n^2}\right),$$

with γ Euler's Constant, 0.5772 . . . , we obtain for the number of comparisons

$$\mu_n \sim 2\{\gamma - 1 + \ln(n+1)\} = 2\ln(n+1) - 0.8456\ldots,$$

$$\sigma_n^2 \sim 2\left\{\gamma + 1 - \frac{2\pi^2}{6} + \ln(n+1)\right\} = 2\ln(n+1) - 3.4253\ldots.$$

Since $|x_\kappa| \leqslant 1$ and $\sigma_n \to \infty$, the central limit theorem applies, yielding

$$P\left(\frac{k_n - \mu_n}{\sigma_n} \leqslant x\right) \to \phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy.$$

Also, by the Berry–Esseen Theorem (cf. Feller [5, p. 544])

$$\sup_x \left| P\left(\frac{k_n - \mu_n}{\sigma_n}\right) - \phi(x) \right| \leqslant 6\frac{r_n}{\sigma_n^3},$$

where

$$r_n = \sum_{\kappa=1}^{n} E|x_\kappa - p_\kappa|^3 = \sum_{\kappa=1}^{n} p_\kappa q_\kappa (p_\kappa^2 + q_\kappa^2).$$

Asymptotically,

$$r_n = -2\left(H_{n+1}^{(4)} - 1\right) + 4\left(H_{n+1}^{(3)} - 1\right) - 3\left(H_{n+1}^{(2)} - 1\right) + H_{n+1} - 1$$
$$\sim -\frac{\pi^4}{45} + 4(0.20205\ldots) - \frac{\pi^2}{2} + \gamma + \ln(n+1) + 1$$
$$= \ln(n+1) - 4.714\ldots,$$

so that the bound

$$6\frac{r_n}{\sigma_n^3} \sim O\left(\frac{1}{\sqrt{\ln(n+1)}}\right),$$

which gives a fairly high rate of convergence to normality for the distribution of the number of comparisons to insert a new symbol. The results given can be used to

analyze a complete binary insertion sort algorithm operating on a random string of symbols. Higher order expansions are also possible for describing this process.

Further, let $M_{n,j}$ be the number of nodes at level j in a random binary tree with n nodes. Since clearly

$$\begin{aligned} &M_{n,j+1} + V_{n,j+1} = 2M_{n,j}, \qquad j = 1, \ldots, n, \quad \text{with} \\ &\qquad M_{n,1} = 1, \quad M_{n,n+1} = 0, \quad n \geqslant 1, \quad \text{whence} \\ &\qquad M_{n,j} = \frac{1}{2} \sum_{r=j}^{n} \left(\frac{1}{2}\right)^{r-j} V_{n,r+1}, \quad \text{or upon taking the expectation} \end{aligned} \tag{4.5}$$

$$E\{M_{n,j}\} = \frac{2^{j-1}}{n!} \sum_{r=j}^{n} \left[\begin{matrix} n \\ r \end{matrix}\right], \qquad j = 1, \ldots, n, \quad n \geqslant 1.$$

(See Appendix, Table A.4.)

Note that

$$E\left\{\frac{1}{2^{j-1}} M_{n,j}\right\} = \frac{1}{n!} \sum_{r=j}^{n} \left[\begin{matrix} n \\ r \end{matrix}\right]$$

is an expected fraction of nodes occupied at level j. It may be called an expected relative thickness of the tree at level j. Looking again at Figure 4.1 we see that we can distinguish between two kinds of vacancies, those which are attached to a leaf (and hence come in pairs) and those attached to an internal node of the tree. We will refer to these two kinds as twin and single vacancies respectively.

Let for $n \geqslant 1, j \geqslant 2$, $V_{n,j}^{(1)}$ be the number of single vacancies and $V_{n,j}^{(2)}$ the number of twin vacancies at level j in a random binary tree with n nodes. Clearly

$$V_{n,j}^{(1)} + V_{n,j}^{(2)} = V_{n,j}; \qquad V_{1,2}^{(1)} = 0, \qquad V_{1,2}^{(2)} = 2.$$

If at time n the values of these variables are

$$v_{n,2}^{(1)}, \ldots, v_{n,n+1}^{(1)} \quad \text{and} \quad v_{n,2}^{(2)}, \ldots, v_{n,n+1}^{(2)},$$

then filling a single vacancy at level j results in

$$V_{n+1,j}^{(1)} = v_{n,j}^{(1)} - 1, \qquad V_{n+1,j+1}^{(2)} = v_{n,j}^{(2)} + 2,$$

with probability $v_{n,j}^{(1)}/(n+1)$, while filling a twin vacancy at level j results in

$$V_{n+1,j}^{(1)} = v_{n,j}^{(1)} + 1, \qquad V_{n+1,j}^{(2)} = v_{n,j}^{(2)} - 2, \qquad V_{n+1,j+1}^{(2)} = v_{n,j+1}^{(2)} + 2,$$

with probability $v_{n,j}^{(2)}/(n+1)$. The remaining vacancy numbers are not changed. If we again introduce the random polynomials

$$W_n^{(1)} = \sum_{j \geqslant 2} V_{n,j}^{(1)} x^{j-1}, \qquad W_n^{(2)} = \sum_{j \geqslant 2} V_{n,j}^{(2)} x^{j-1},$$

we obtain for their expected values the equations

$$E\{W_{n+1}^{(1)}(x)\} = \frac{n}{n+1} E\{W_n^{(1)}(x)\} + \frac{1}{n+1} E\{W_n^{(2)}(x)\},$$

$$E\{W_{n+1}^{(2)}(x)\} = \frac{2x}{n+1} E\{W_n^{(1)}(x)\} + \frac{n-1+2x}{n+1} E\{W_n^{(2)}(x)\},$$

$$n \geqslant 1, \qquad \text{with} \quad E\{W_1^{(1)}(x)\} = 0, \quad E\{W_1^{(2)}(x)\} = 2x.$$

Since

$$E\{W_n^{(1)}(x)\} + E\{W_n^{(2)}(x)\} = E\{W_n(x)\}$$

we obtain by substitution

$$E\{W_{n+1}^{(1)}(x)\} = \frac{n-1}{n+1} E\{W_n^{(1)}(x)\} + \frac{1}{n+1} E\{W_n(x)\},$$

from which by using (4.2) we have

$$\begin{aligned} E\{W_n^{(1)}(x)\} &= \frac{1}{n(n-1)} \sum_{r=1}^{n-1} rE\{W_r(x)\} \\ &= \frac{1}{n(n-1)} \sum_{r=1}^{n-1} \frac{1}{(r-1)!} \sum_{j=2}^{r+1} \left[{r \atop j-1} \right] (2x)^{j-1} \\ &= \frac{1}{n(n+1)} \sum_{j=2}^{n} (2x)^{j-1} \sum_{r=j-1}^{n-1} \frac{1}{(r-1)!} \left[{r \atop j-1} \right], \qquad n > 1. \end{aligned} \tag{4.6}$$

Thus,

$$E\{V_{nj}^{(1)}\} = \frac{2^{j-1}}{n(n+1)} \sum_{r=j-1}^{n-1} \frac{1}{(r-1)!} \left[{r \atop j-1} \right], \tag{4.7}$$

$j = 2, \ldots, n$; $n > 1$, while the corresponding expression for twin vacancies is obtained by subtracting (4.7) from (4.3) with (4.7) set equal to zero for $j = n + 1$ (see Appendix, Table A.6).

We conclude this section by computing the expected number of leaves in a random binary tree with n nodes. From the first equality in (4.6) the expected number of all single vacancies equals

$$E\{W_n^{(1)}(1)\} = \frac{1}{n(n-1)} \sum_{r=1}^{n-1} rE\{W_r(1)\} = \frac{1}{n(n-1)} \sum_{r=1}^{n-1} r(r+1),$$

since by (4.2) $E\{W_n(1)\} = n + 1$. The latter sum equals $\frac{1}{3}(n+1)$ so that the expected number of all twin vacancies is $n + 1 - \frac{1}{3}(n+1)$. The expected number of leaves is clearly half the number of twin vacancies, that is $\frac{1}{3}(n+1)$. Thus, in a random binary tree on the average about $\frac{1}{3}$ of the nodes are leaves.

REMARK. Having in mind the process of growing random binary trees as described above we can also look at a random process

$$N_k = \min\{n : H_n > k\}, \qquad N_0 = 1, \tag{4.8}$$

i.e., the time (= number of symbols) needed to grow a tree over the height k (see Appendix, Table A.8). From (4.8) clearly $P(N_k > n) = P(H_n \leq k) = F(n,k)$ so that $g_k(x) = 1 - (1-x)f_k(x)$ is the ordinary probability generating function $g_k(x) = \sum_{n\geq 1} P(N_k = n)x^n$ of the random variable N_k. Denoting

$$\tilde{\mu}(m,k) = E\{N_k(N_k-1)\cdots(N_k-m+1)\}$$

the mth factorial moment of N_k and using the fact that

$$\tilde{\mu}(m,k) = \frac{d^{(m)}}{dx^m} g_k(1) = m \frac{d^{(m-1)}}{dx^{m-1}} f_k(1)$$

we obtain by applying Leibnitz formula to (3.13) the relation

$$\tilde{\mu}(m+2,k+1) = \frac{1}{m+1}\sum_{j=0}^{m}\binom{m+2}{j+1}\tilde{\mu}(j+1,k)\tilde{\mu}(m-j+1,k),$$

$m \geqslant 0$, $k \geqslant 0$. In particular with $m = 0$ this becomes

$$E\{N_{k+1}(N_{k+1}-1)\} = 2(E\{N_k\})^2, \qquad k \geqslant 0, \quad \text{or}$$

$$\operatorname{Var}\{N_{k+1}\} = 2(E\{N_k\})^2 + E\{N_{k+1}\}\left[1 - E\{N_{k+1}\}\right].$$

Unfortunately, it is the first moment $E\{N_k\}$, which is hard to obtain for large k (see Appendix, Table A.7).

Since completion of this work, some related material has been published by Ruskey [11].

5. Acknowledgment. We are deeply indebted to Professor Herbert Rutemiller for his early contributions and encouragement, without which we would never have continued our efforts.

Appendix

TABLE A.1

Binary Trees with n Nodes and Height k, t(n, k) (2.3)

No. of Symbols	n!	Height 0	1	2	3	4	5	6	7	8	9	10
0	1	1										
1	1	0	1									
2	2	0	0	2								
3	5	0	0	1	4							
4	14	0	0	0	6	8						
5	42	0	0	0	6	20	16					
6	132	0	0	0	4	40	56	32				
7	429	0	0	0	1	68	152	144	64			
8	1430	0	0	0	0	94	376	480	352	128		
9	4862	0	0	0	0	114	844	1440	1376	832	256	
10	16796	0	0	0	0	116	1744	4056	4736	3712	1920	512

TABLE A.2

Partitions of Permutations of n Symbols into Binary Trees of Height k (Adapted from (3.9))

No. of Symbols	n!	Height 0	1	2	3	4	5	6	7	8	9	10
0	1	1										
1	1	0	1									
2	2	0	0	2								
3	6	0	0	2	4							
4	24	0	0	0	16	8						
5	120	0	0	0	40	64	16					
6	720	0	0	0	80	400	208	32				
7	5040	0	0	0	80	2240	2048	608	64			
8	40320	0	0	0	0	11360	18816	8352	1664	128		
9	362880	0	0	0	0	55040	168768	104448	30016	4352	256	
10	3628800	0	0	0	0	253440	1508032	1277568	479040	99200	11008	512

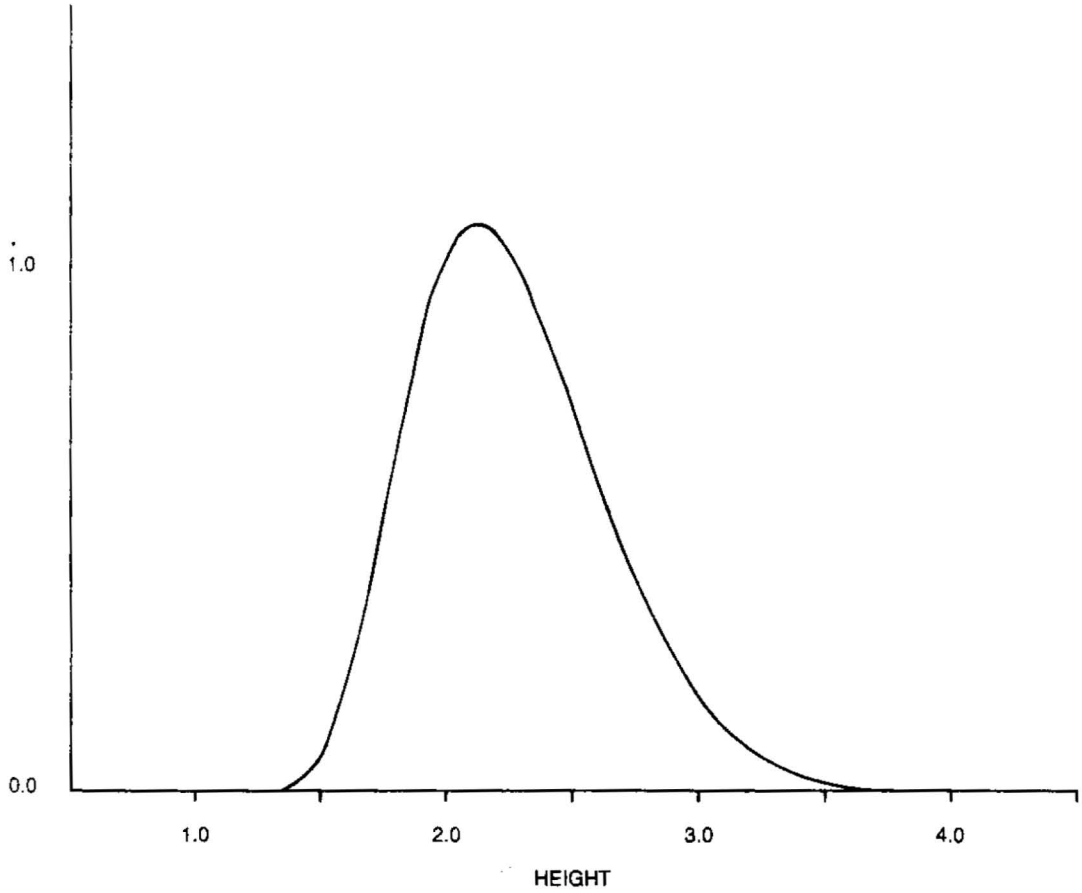

FIGURE A.1. Normalized asymptotic distribution of binary tree height (2.42).

TABLE A.3

Random Binary Tree Height (Adapted from (3.9) and (3.23))

No. of Symbols	Height		
	Minimum	Mean	Variance
0	0	0.000	0.000
1	1	1.000	0.000
2	2	2.000	0.000
3	2	2.667	0.222
4	3	3.333	0.222
5	3	3.800	0.426
6	3	4.267	0.507
7	3	4.670	0.570
8	4	5.018	0.682
9	4	5.340	0.774
10	4	5.641	0.837
50	6	10.810	2.051
100	7	13.286	2.522
150	8	14.778	2.772
200	8	15.852	2.934
250	8	16.693	3.052
300	9	17.385	3.147
$\sim n$	$\lceil$[1]$1.443 \ln(n)$	$\sim 3.259 \ln(n)$	

[1]$\lceil$ indicates next higher integer.

TABLE A.4

Expected Nodes by Level in Random Binary Trees (4.5)

No. of Symbols	Level 1	2	3	4	5	6	7	8	9	10
0	0.0000									
1	1.0000									
2	1.0000	1.0000								
3	1.0000	1.3333	0.6667							
4	1.0000	1.5000	1.1667	0.3333						
5	1.0000	1.6000	1.5333	0.7333	0.1333					
6	1.0000	1.6667	1.8111	1.1222	0.3556	0.0444				
7	1.0000	1.7143	2.0286	1.4794	0.6254	0.1397	0.0127			
8	1.0000	1.7500	2.2036	1.8016	0.9171	0.2786	0.0460	0.0032		
9	1.0000	1.7778	2.3476	2.0911	1.2155	0.4514	0.1028	0.0131	0.0007	
10	1.0000	1.8000	2.4684	2.3515	1.5122	0.6494	0.1828	0.0323	0.0032	0.0001

TABLE A.5

Expected Vacancies by Level in Random Binary Trees (4.3)

No. of Symbols	Total Vacancies	Level 1	2	3	4	5	6	7	8	9	10	11
0	1	1.0000										
1	2	0.0000	2.0000									
2	3	0.0000	1.0000	2.0000								
3	4	0.0000	0.6667	2.0000	1.3333							
4	5	0.0000	0.5000	1.8333	2.0000	0.6667						
5	6	0.0000	0.4000	1.6667	2.3333	1.3333	0.2667					
6	7	0.0000	0.3333	1.5222	2.5000	1.8889	0.6667	0.0889				
7	8	0.0000	0.2857	1.4000	2.5778	2.3333	1.1111	0.2667	0.0254			
8	9	0.0000	0.2500	1.2964	2.6056	2.6861	1.5556	0.5111	0.0889	0.0063		
9	10	0.0000	0.2222	1.2079	2.6041	2.9667	1.9796	0.8000	0.1926	0.0254	0.0014	
10	11	0.0000	0.2000	1.1316	2.5853	3.1908	2.3750	1.1159	0.3333	0.0614	0.0063	0.0003

TABLE A.6

Expected Single Vacancies by Level in Random Binary Trees (4.7)

No. of Symbols	Level 1	2	3	4	5	6	7	8	9	10
0	1.0000									
1	0.0000									
2	0.0000	0.1667								
3	0.0000	0.1111	0.1111							
4	0.0000	0.0750	0.1167	0.1000						
5	0.0000	0.0533	0.1022	0.1467	0.1067					
6	0.0000	0.0397	0.0862	0.1603	0.2032	0.1270				
7	0.0000	0.0306	0.0724	0.1585	0.2680	0.2993	0.1633			
8	0.0000	0.0243	0.0612	0.1501	0.3057	0.4643	0.4603	0.2222		
9	0.0000	0.0198	0.0522	0.1394	0.3241	0.6019	0.8226	0.7309	0.3160	
10	0.0000	0.0164	0.0449	0.1283	0.3299	0.7084	1.1967	1.4804	1.1895	0.4655

TABLE A.7

Distribution of Comparisons for Insertion of Random Symbol in Random Binary Tree ((4.4), Expected Fraction of Vacancies by Level in Random Binary Trees with κ = Level − 1)

No. of Symbols	Comparisons, κ										
	0	1	2	3	4	5	6	7	8	9	10
0	1.0000										
1	0.0000	1.0000									
2	0.0000	0.3333	0.6667								
3	0.0000	0.1667	0.5000	0.3333							
4	0.0000	0.1000	0.3667	0.4000	0.1333						
5	0.0000	0.0667	0.2778	0.3889	0.2222	0.0444					
6	0.0000	0.0476	0.2175	0.3571	0.2698	0.0952	0.0127				
7	0.0000	0.0357	0.1750	0.3222	0.2917	0.1389	0.0333	0.0032			
8	0.0000	0.0278	0.1440	0.2895	0.2985	0.1728	0.0568	0.0099	0.0007		
9	0.0000	0.0222	0.1208	0.2604	0.2967	0.1980	0.0800	0.0193	0.0025	0.0001	
10	0.0000	0.0182	0.1029	0.2350	0.2901	0.2159	0.1014	0.0303	0.0056	0.0006	0.0000

TABLE A.8

Number of Random Symbols Needed to Grow a Binary Tree of Height Exceeding k (4.8)

Height k	Number of Symbols		
	$E\{N_k\}$	$\sigma\{N_k\}$	Range
0	1	0	1
1	2	0	2
2	3.33	0.491	3–4
3	5.13	0.995	4–8
4	7.53	1.861	5–16
5	10.75	2.931	6–32
6	15.02	4.533	7–64
7	20.67	6.680	8–128
8	28.13	9.557	9–256
9	37.92	13.512	10–512

References

[1] Abramowitz, M. and Stegun, I., eds. (1964), *Handbook of Mathematical Functions*, National Bureau of Standards Applied Mathematics Series, No. 55, Washington, D.C.

[2] Brown, G. and Rutemiller, H. (1973). On the Probability Distribution of Tree Depth for Randomly Grown Binary Trees (unpublished manuscript).

[3] ——— and Shubert, B. (June 1976). On Random Binary Trees. Naval Postgraduate School Technical Report *NPS55Bw76061*, 46 pp.

[4] du Bruijn, N. (1961). *Asymptotic Methods in Analysis*. North-Holland. Amsterdam.

[5] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications, Volume II*. John Wiley and Sons, New York.

[6] Jolley, L. (1961). *Summation of Series*. Dover, New York.

[7] Knuth, D. (1973). *The Art of Computer Programming, Vol. 1, Fundamental Algorithms*. Addison Wesley, Reading, Massachusetts.

[8] ------. (1973). *The Art of Computer Programming, Vol. 3, Sorting and Searching*. Addison Wesley, Reading, Massachusetts.

[9] Renyi, A. and Szekeres, G. (1967). On the Height of Trees. *Austral. Math. Soc. J.* 7 497.

[10] Robson, J. (1979). The Height of Binary Search Trees. *Austral. Comput. J.* **11** 151.

[11] Ruskey, F. (1980). On The Average Shape of Binary Trees. *SIAM J. Algebraic Discrete Methods* 1 43.

[12] Szekeres, G. (1958). Regular Iteration of Real and Complex Functions. *Acta Math.* **100**.

NAVAL POSTGRADUATE SCHOOL, MONTEREY, CALIFORNIA 93940